Transformer-based Dual Relation Graph for Multi-label Image Recognition

Supplementary Material

This Appendix provides detailed parametric studies and explainable visualization of our proposed approach. Appendix A analyzes the effect of contextual information in our motivation. Then we report the AP of each category on MS-COCO dataset in Appendix B. Appendix C provides detailed parametric studies of different modules. Lastly, we provide more visualization results in Appendix D.

A. Explainable Structural Relation

We discuss the effects of structural relation in our motivation, here we conduct detailed visualized results in Fig. 1 to articulate this point of view.

Challenge of semantic confusion. The ground-truth labels of source image in Fig. 1 are {*boat, person, snowboard*}, while our baseline model could not distinguish *skateboard* and *snowboard* due to their **similar appearance**. In addition, considering the **co-occurrence** of multiple labels, {*person, snowboard*} and {*person, skateboard*} are both high frequency collocations, which could not be directly solved by conventional GCN-based methods aforementioned in the main manuscript. These two unsolved problems lead to a burning challenge of semantic confusions in multi-label recognition.

Contextual information for recognition. However, we human being can easily recognize the object as a *snowboard* rather than a *skateboard*, due to the context *snow* and the interaction with *person*. Towards this ends, we introduce Transformer architecture [5] to capture long-term contextual information and build position-wise relationships between different objects to build structural relation. As shown in Fig. 1, our structural relation graph module could effectively capture the context *snow* and identify the object as a *snowboard* with certainty. Especially, the response map of *snowboard* shows a obvious structural relation between *snowboard* and context *snow* and the response map *boat* also shows a obvious structural relation between *boat* and context *water* around.

Moreover, for individual objects such as *person*, our proposed network has the potential to understand the high-level information while focusing more on the compact regions, while the baseline model introduces more noisy backgrounds.

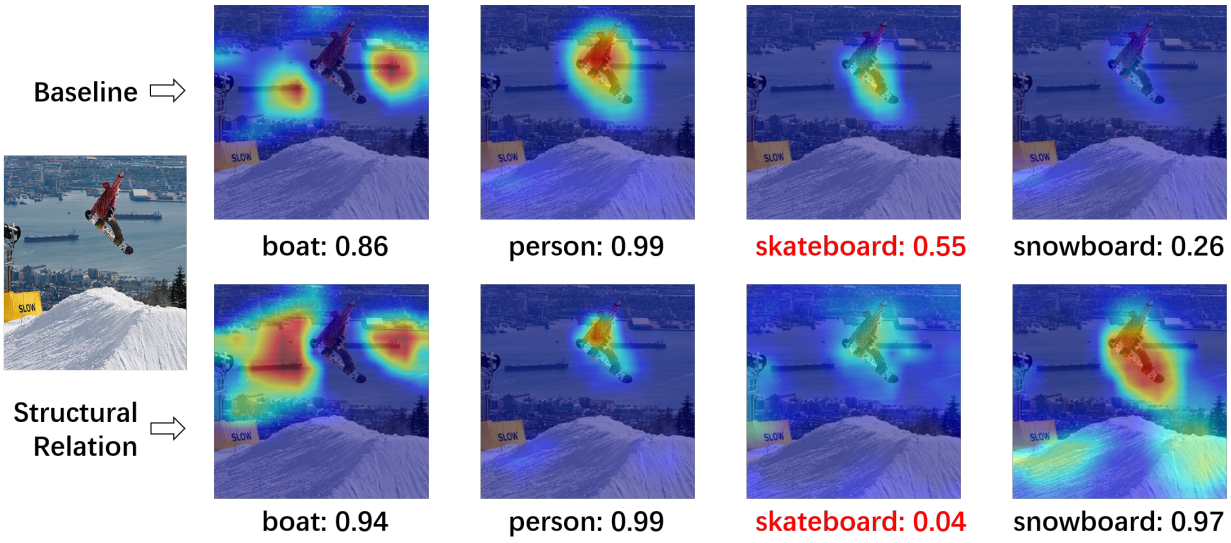

Figure 1. A representative case of the proposed structural relation. Our structural relation graph module could capture long-term contextual information and easily distinguish the objects with similar appearance, *e.g.*, *skateboard* and *snowboard*.

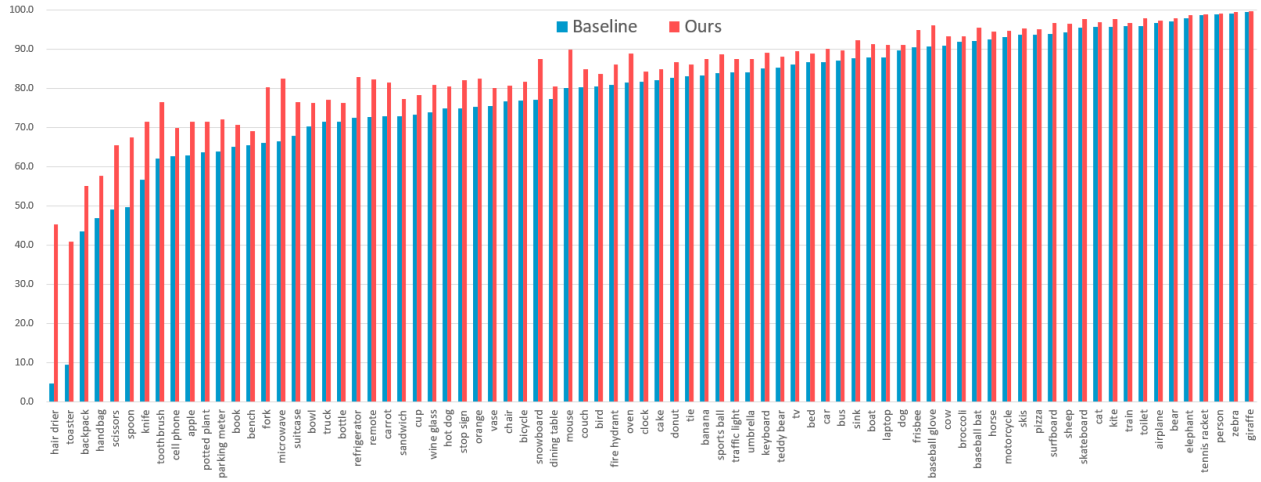

Figure 2. AP of each category of baseline model and our proposed approach on MS-COCO dataset. Our approach has significant improvement on almost all categories, especially for small object categories, *e.g.*, *hair drier* and *scissors*, which demonstrates the strong capacity of our model in capturing small objects.

B. Comparison results on MS-COCO

As shown in Fig. 2, we exhibit the comparison chart of AP performance of each category on MS-COCO. Our proposed approach improves our baseline ResNet-101 from 78.6% to 84.6% (6.0% higher) in terms of mAP. It is obvious that our approach has significant improvement on almost all categories, especially for categories with small scales, *e.g.*, *hair drier* and *toaster*, which demonstrates the strong capacity of our approach in capturing small objects. Meanwhile, our approach shows the potential in distinguishing objects with visually similar appearances, *e.g.*, *backpack* and *handbag*.

C. Parametric Study

C.1. Weights of structural and semantic relation

As mentioned in Eq. (3) of the main manuscript, the structural relation graph and the semantic relation graph are weighted fused to obtain the final prediction. We apply a weight coefficient α on structural relation graph module and $(1-\alpha)$ on semantic relation graph module. As shown in Fig. 3, we set $\alpha = 0.7$ to achieve the best performance 84.6% on MS-COCO and 95.0% on VOC 2007 dataset.

C.2. Hidden dimension of Transformer and GCN

The hidden dimension of Transformer and GCN are two important factors in our experiments. We conduct detailed parametric studies on different combinations of dimensions on MS-COCO in Tab. 1. In experiments, we set $C_T = 512$ and $C_G = 512$ to achieve the best performance.

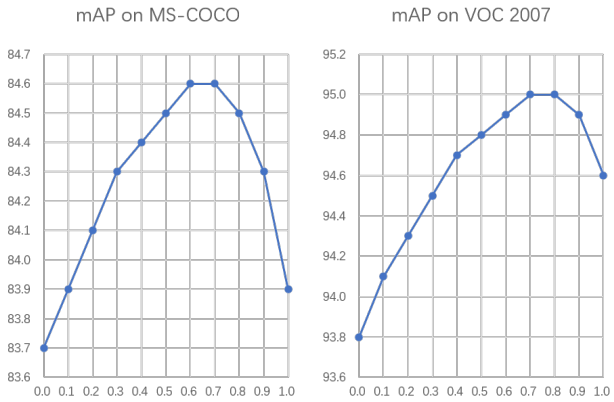

Figure 3. The performance of the final prediction with different weight coefficients α on MS-COCO and VOC 2007.

Table 1. Parametric study of the hidden dimension of Transformer and GCN on MS-COCO.

C_T	256	512	512	512	512	1024
C_G	512	256	512	1024	2048	1024
mAP	84.4	84.4	**84.6**	84.4	84.4	84.5

C.3. Multi-head Self-attention in Transformer

As two important factors in Transformer units, we conduct detailed parametric studies of the number of encoder layers n and attention heads m in Tab. 2. It can be found that more attention heads could lead to better performance. However, more attention heads also introduce more computation cost, hence we set m as 4 rather than 8 as a trade-off between performance and computation cost. Then we explore different numbers of encoder layers from 1 to 6 and

set n as 3 to obtain the best performance.

Table 2. Parametric study of the number of encoder layers n and attention heads m on MS-COCO.

n	1	2	3	4	5	6	3	3	3
m	4	4	4	4	4	4	1	2	8
mAP	84.4	84.4	**84.6**	84.4	84.4	84.5	84.5	84.5	**84.6**

C.4. Position Encoding

We introduce position encoding to retain the spatial structure information following [1]. As shown in Tab. 3, the result without positional information can be found in the first row. By incorporating the absolute positional encoding operation (sinusoid encoding function in [5]), the performance of structural relation module slightly drops. In this paper, we adopt relative positional encoding with learned encoding parameters to encode the unique positional information, which slightly improves the joint performance from 84.4% to 84.6%.

Table 3. Parametric study of position encoding manner on MS-COCO.

Position Encoding	$\mathcal{R}_{Structural}$	$\mathcal{R}_{Joint}$
None	83.8	84.4
Absolute Encoding	83.7	84.4
Relative Encoding	**83.9**	**84.6**

C.5. Numbers of GCN layers

We report the performance results with different numbers of GCN layers in our semantic relation graph module in Tab. 4. We notice that the performance slightly drops when the number of GCN layers increases. Hence we set the number of GCN layers as 1 to obtain the best recognition performance.

Table 4. Parametric study of the number of GCN layers on MS-COCO.

GCN layers	$\mathcal{R}_{Semantic}$	$\mathcal{R}_{Joint}$
1	**83.7**	**84.6**
2	83.5	84.5
3	83.4	84.5

C.6. Pooling operation of Semantic-aware Constraints

As mentioned in the main manuscript, we utilize the top-k max-pooling with a threshold of 5% to squeeze the high-response spatial information for semantic-aware constraints. We explore different global pooling operations on MS-COCO in Tab. 5, it can be found that focusing on the high-response area of each channel by top-k max-pooling achieves better performance than focusing on the whole area by global average-pooling or only one point by global max-pooling. For top-k max-pooling, we explore different thresholds in Tab. 5 and set the threshold as 5% to obtain the best performance.

Table 5. Ablation study of different global pooling operations in semantic-aware constraints on MS-COCO dataset.

Pooling Method	GAP	GMP	$KMP_{5\%}$	$KMP_{10\%}$	$KMP_{20\%}$	$KMP_{50\%}$
mAP	84.4	84.3	**84.6**	84.4	84.5	84.4

D. Visualization Results

In this section, we visualize the Class Activation Maps by Grad-CAM [4] on two popular datasets, *i.e.* MS-COCO [3] and VOC 2007 [2] dataset.

D.1. VOC 2007 dataset

VOC 2007 is a 'easier' dataset compared to MS-COCO. We provide the visualization results on VOC 2007 in Fig. 4. It can be found that our structural relation could obtain more accurate localization and higher confidence with global contextual information, *e.g. bottle* in Fig. 4 a) and {*pottedplant, sofa*} in Fig. 4 b).

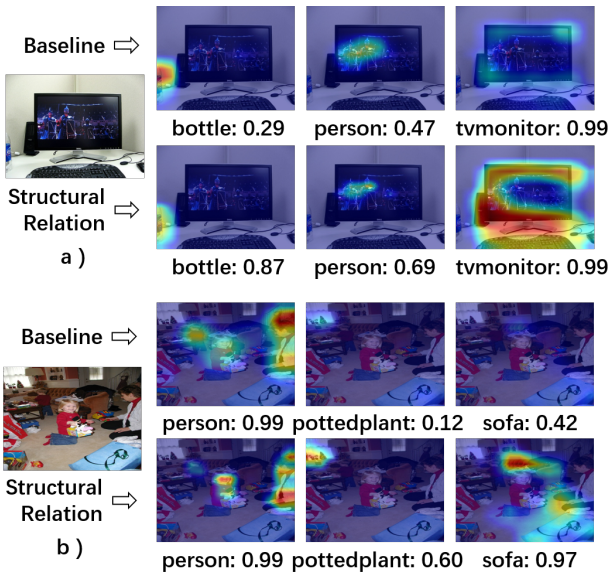

Figure 4. Visualization analyses of baseline and our proposed structural relation graph module on VOC 2007. We present several labels for demonstration.

D.2. MS-COCO dataset

MS-COCO is a challenging dataset for multi-label recognition tasks due to several characteristics, *e.g.*, small

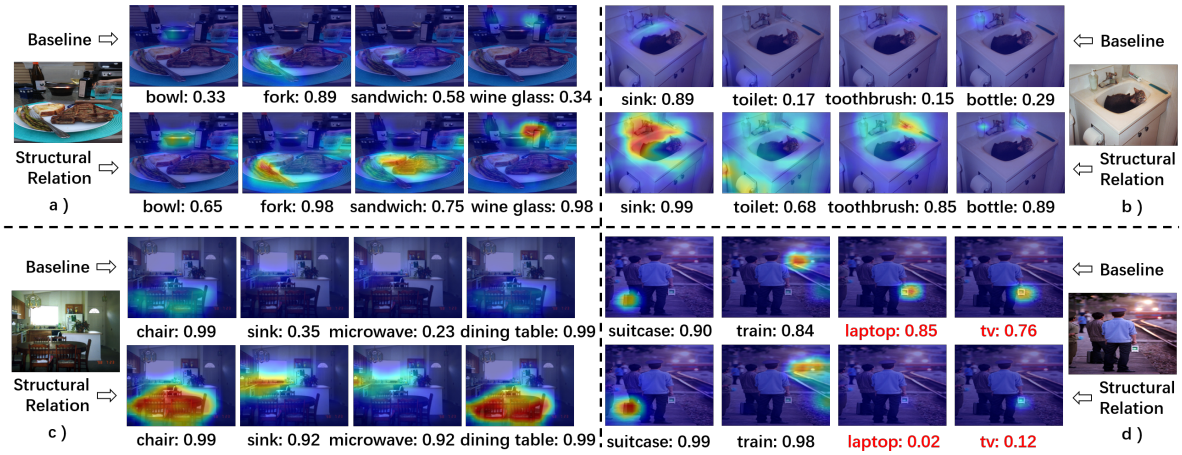

Figure 5. Visualization analyses of baseline and our proposed structural relation graph module on MS-COCO. We present several labels for demonstration and the labels not presented in the image are highlighted in red.

objects and complex scene. It can be found in our main manuscripts that our Transformer-based structural relation module provides better results and localization than baseline. More results can be found in Fig. 5, our structural relation could not only figure out the **small objects**, *e.g. wine glass* in Fig. 5 a) and *toothbrush* in Fig. 5 b), but also distinguish the objects in **complex scenes**, *e.g.* {*sink, microwave*} in Fig. 5 c). Besides, our approach has strong capacity to distinguish the objects with **similar appearance**, *e.g.*, *laptop* and *tv* in Fig. 5 d).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 1, 3